연어 정보를 이용한 한국어 철자 검사기의 기능 개선*

심철민, 김현진, 김영진, 권혁철
부산대학교 전자계산학과

Improvement of a Korean Speller with Collocation of Parts of Speech

Chul-Min Sim, Hyun-Jin Kim, Young-Jin Kim, Hyuk-Chul Kwon
Departmemt of Computer Science, Pusan National University

요 약

본 논문에서는 한 어절 단위에서 다수 어절 단위로 그 고려 영역을 확장한 개선된 철자 검사기를 제시한다. 개선된 철자 검사기는 1) 한 어절 철자 검사 교정부, 2) 연어 규칙 처리부, 3) 문장 부호 규칙 처리부로 구성된다. 한 어절 철자 검사 교정부는 기존의 철자 검사기와 같은 기능을 수행한다. 연어 규칙 처리부는 형태소 간의 연어 관계를 이용하여 7가지로 유형 분류된 어절 간 오류를 처리한다. 문장 부호 처리부는 문장 부호 자체의 오류와 문장 부호을 참조하여 좌우 어절들의 오류를 검사한다. 현재 256가지의 연어 규칙과 51가지의 문장 부호 규칙이 구축되어 있다.

본 논문에서 제시한 개선된 철자 검사기는 한국어 문체 검사기(Korean Style Checker) 로서 의의를 가지며, 형태소의 연어 정보는 향후 파싱 등의 문장 분석이나 의미 분석에 중요한 자료로 이용될 수 있을 것으로 기대된다.

I.서론

기존의 한국어 철자 검사기는 맞춤법 오류, 띄어쓰기 오류, 사투리 및 순화 대상 용어 사용 등의 한 어절 단위 철자법 오류를 대상으로 하고 있다[1,2,3]. 그러나 문서 내에 존재하는 오류는 이러한 오류들 외에도 문장 부호 오류, 문맥상 잘못된 어휘 사용, 높임말 사용 오류, 시제 오류 등 다양하다. 특히 이 중에서도 문장 부호 오류와 문맥상 잘못된 어휘 사용 등은 띄어쓰기 오류, 맞춤법 오류 다음으로 발생 빈도가 높다[4].

본 논문에서는 한 어절 단위의 분석 영역을 다수 어절로 확장한 개선된 철자 검사기를 제안한다. 개선된 철자 검사기는 1) 한 어절 단위의 철자 검사 교정부, 2) 연어 규칙 검사부, 3) 문장 부호 규칙 검사부로 구성된다. 한 어절 철자 검사 교정부는 한 어절 내에서의 맞춤법 오류와 띄어쓰기 오류를 검사한다. 연어 규칙 검사부에서는 다수 어절을 참조하면서 지식 베이스로 구현된 연어 제약 조건을 검사한다. 문장 부호 규칙 검사부는 다수 어절을 참조하면서 맞춤법 표준안에 기반한 문장 부호별 처리 규칙을 적용한다[4,5,6].

연어 규칙 검사부에서 참조하는 연어 정보는 오류의 원인에 따라 발음의 유사성에 의한 오류, 의미의 유사성에 의한 오류, 명사 간의 결합 오류, 띄어쓰기 오류, 어미/조사의 오류, 부사-서술어 오류의 6가지로 세분되며, 이의 처리 방법은 붙여쓰기, 띄어쓰기, 한 단어 대치, 다수 단어 대치, 복합명사 구성 제약, 좌우 접속 제약 그리고 특수 유형의 7가지로 구분된다.

현재 총 256가지의 연어 정보와 51가지의 문장 부호 규칙이 지식 베이스로 구축되어 있다. 이러한 오류 규칙들이 충분히 구축된다면 다수 어절을 고려한 철자 검사기의 검사와 교정 성능이 동시에 개선될 것이며, 오류 검사를 하기 위해 구축된 연어 정보는 향후 파싱 등의 문장 분석에 이용될 수 있다.

II.형태소 간 연어 정보의 분류 및 처리

본 논문에서 대상으로 하는 다수 어절 단위 오류는 한 어절 단위의 처리로는 오류를 판단할 수 없이 올바른 어절로 처리되므로, 좌우 어절들의 정보를 이용하여야만 검사하고 교정할 수 있는 오류를 의미한다.

*본 논문은 STEP2000('94핵심S/W 기술개발 사업) 중 '국어정보처리 기술 개발' 사업의 과제인 한국과학기술원 지능형 처리기 개발 과제 연구비에 의해 연구되었음.

[예문-a] 옷에 땀이 베어서 입을 수가 없다. (X)
[예문-b] 옷에 땀이 배어서 입을 수가 없다. (O)

[예문-a]는 철자법상으로는 맞는 문장이다. 그러나 '땀'이라는 단어는 '물기가 스미어 젖다'의 의미가 있는 '배다'와 연관 관계가 있으므로 오류를 인식할 수 있게 된다. 이와 같이 철자법이 올바른 어절 중에서도 의미적 오류로 인해서 교정이 필요한 경우가 빈번히 발생하고 있다. 그리고 철자법 오류의 경우 기존의 철자 검사기가 사용한 음소 대치나 음절 대치 등은 교정 시간이 많이 걸리고 부정확한 경우가 많이 발생한다. 이를 해결하기 위해 일반인들이 빈번히 범하는 오류 단어에 대해서 형태소 간의 연어 정보를 이용하여 오류를 추정하고 교정하는 기법을 사용할 수 있다.

본 논문에서는 빈번히 발생하는 의미적인 오류를 좌우의 연어 관계를 이용하여 교정할 수 있는 연어 관계에 기반한 개선된 철자 검사기를 제안한다. 연어 정보의 구축을 하기 위해 사용자들이 범하는 오류의 유형을 분석하고, 그 처리 방안을 제시한다.

2.1 오류 유형의 발생 원인별 분류

어절 간의 오류는 그 발생 원인에 따라 발음의 유사성에 의한 오류, 의미의 유사성에 의한 오류, 명사 간의 결합 오류, 띄어쓰기 오류, 어미/조사의 오류, 부사-서술어 오류로 나누어진다.

발생 원인별 분류	예문 및 교정 결과	조사된 규칙수
발음의 유사성에 의한 오류	편지를 붙이고 오는 길이다.(X) 편지를 부치고 오는 길이다.(O)	122 개
의미 유사성에 의한 오류	이번 정기의 성패는 너에게 달려 있다.(X) 이번 정기의 승패는 너에게 달려 있다.(O)	50 개
명사 간 결합오류	한국일생 (X) 이미니인생 (O)	151 개
띄어쓰기 오류	날 사랑하는 이는 이미니 밖에 없다.(X) 난 사랑하는 이는 이미니밖에 없다.(O)	27 개
어미/조사의 오류	밥은 먹었는가 모르겠네 (X) 밥은 먹었는지 모르겠네 (O)	57 개
부사-서술어 오류	너는 결코 그것을 할 수 있다. (X) 너는 결코 그것을 할 수 있다. (O)	15 개

[표1] 어절 간 오류의 발생 원인별 분류

발음 유사성에 의한 오류는 의미는 서로 다르나 발음이 비슷해서 일반인들이 범하게 되는 오류이다. 발음 유사성을 가진 단어 쌍을 구하기 위해 전자사전 내의 단어들 중 두 음소까지 차이나는 단어의 쌍을 구하였고, 그 중 발음 규칙상 발음이 같은 단어 쌍을 추출하여 수작업 검증을 통해 구축하였다. 의미 유사성에 의한 오류는 의미는 비슷하나 상황에 따라 쓰이는 용도가 다른 단어 쌍에 대한 오류이다. 명사 간의 결합 오류는 의미적으로 연결될 수 없는 명사끼리 복합 명사를 구성했을 경우의 오류이다. 띄어쓰기 오류는 좌우 상황에 따라 붙여 써야 할 경우도 있고 띄어 써야 하는 경우도 있는 단어에 대한 오류이다. 어미/조사의 오류는 용언이나 명사의 종류에 따라 어미나 조사를 다르게 써야하는 경우의 오류이다. 부사-서술어 오류는 부사와 서술어 간의 호응 관계에 대한 오류이다.

2.2 오류 유형의 처리 방법

본 논문에서는 발생 원인에 따라 조사된 어절 간 오류 유형들을 형태소 간의 연어 관계에 기반하여 처리한다. 이를 위해서는 어절 간 오류 유형들을 연어 관계를 참조하는 형태와 적용되는 교정 기법에 따라 재구성한다. 어절 간 오류는 처리 방법에 따라서 붙여쓰기, 띄어쓰기, 한 단어 대치, 다수 단어 대치, 복합 명사 제약, 좌우 접속 제약, 특수 유형으로 구분된다. [표 2]는 어절 간 오류의 처리 방법 별 분류 결과이다. 현재 이 중 256가지의 규칙이 지식 베이스로 구축되있으며, 나머지 규칙들도 계속 추가해 나갈 계획이다.

처리방법에 따른 분류	예문 및 교정 결과	조사된 규칙수
붙여 쓰기	음식을 손으로 먹지 않은 체 하고 있었다.(X) 음식을 손으로 먹지 않은 체하고 있었다. (O)	27 개
띄어 쓰기	모두다 모여 주세요. (X) 모두 다 모여 주세요. (O)	15 개
한 어절 대치	바다에 흘린 기름이 엉켜서 지저분해졌다. (X) 바다에 흘린 기름이 엉겨서 지저분해졌다. (O)	160 개
다수 어절 대치	색종이를 팔라 부치고 있었다. (X) 색종이를 잘라 붙이고 있었다. (O)	20 개
복합 명사 제약	처리가수 (X) 인기가수 (O)	151 개
좌우 접속 제약	너는 그것을 반드시 하고 있다. (X) 너는 그것을 반드시 해야 한다. (O)	35 개
특수 유형	학교를 가는 아이들이 보였다. (X) 학교에 가는 아이들이 보였다. (O)	14 개

[표2] 처리 방법에 따른 분류

철자 검사 교정부에서의 붙여쓰기 교정은 명사 뒤의 조사는 무조건 붙여 보는 등의 일반적인 규칙을 적용하지만 연어 처리부의 붙여쓰기는 이외에 좌우 어절의 연관 정보의 분석을 필요로 하는 오류를 처리한다.

띄어쓰기 교정은 의존명사나 부사로도 쓰일 수 있고 조사로도 쓰일 수 있는 단어에 대하여 앞 뒤 연어 정보를 참조하여 띄어 쓸 지 붙여쓸 지를 결정한다. 한 단어 대치 교정은 가장 많은 유형으로써 좌우 어절의 정보가 현재 어절의 연어 오류(anti-collocation)조건에 일치할 경우를 처리한다[7].

다수 단어 대치 교정은 다수 어절에 걸친 연어 오류 조건이 일치할 경우를 처리한다. 복합 명사 제약은 오류 원인별 분류에서와 같이 의미적으로 연관성이 없는 명사 간의 결합을 방지한다. 특수 유형은 실제 교정되어야 할 위치가 아니라 그 앞 뒤에서 연어정보를 검사하는 경우이다. 이러한 처리 방법별 오류 유형은 연어 관계 지식 베이스로 구축된다.

III.문장 부호 오류의 처리

문장 부호는 마침표, 쉼표, 따옴표, 묶음표, 이음표, 드러냄표, 안드러냄표 등 일곱 개의 범주로 나뉜다. 한글 맞춤법 표준안에는 총 24개의 문장 부호와 58가지의 문장부호 사용 규칙이 명시되어 있다[5].

본 논문에서는 문장 부호의 처리 범주를 크게 두 가지로 나눈다.

첫째, 문장 부호 자체의 오류에 대해 검사 교정한다. 문장 부호의 사용 오류에는 두 가지가 있다. 문장 부호를 사용규칙에 따르지 않고 잘못 사용한 경우와 문장 부호를 써야 할 위치에 사용하지 않은 경우가 이에 해당한다.

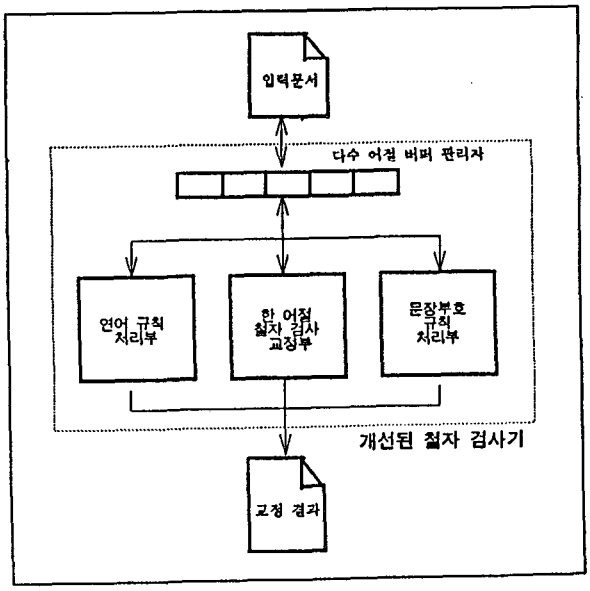
둘째, 문장 부호를 참조하여 좌우 어절들의 오류 여부를 검사하고 오류를 교정한다. 이 처리를 통해 문장 부호 검사부는 문장 부호 자체의 검사 및 교정 뿐만 아니라, 좌우의 어절들까지 검사하므로 철자 검사기의 처리 영역을 더욱 확장시킨다.

문장 부호 처리부가 위의 두 범주를 처리하는 방식으로는 두 가지가 있다. 문장 부호의 앞,뒤 두 어절 정도를 분석하여 처리하는 방법과 한 문장 이상의 분석을 통해 처리하는 방법이다. 전자의 방법으로 처리 가능한 예로 [표 3]-(1), [표 3]-(3), [표 3]-(4)등의 경우가 있다. 반면 한 문장 이상의 분석이 필요한 경우가 [표 3]-(2)이다. 그 외에도 따옴표나 묶음표 등과 같이 여는 부호와 닫는 부호의 쌍의 일치 여부를 검사하기 위해서는 한 문장 단위 이상의 처리가 필요하다.

처리 범주	예문 및 교정 결과	처리 규칙수
문장부호 자체의 사용 오류	1. 1995. 7. 24 (X) → 1995. 7. 24. (O) 2. "침착해. 하늘이 무너져도 솟아날 구멍이 있다고 하니까."고 그가 말했다. (X) → "침착해. '하늘이 무너져도 솟아날 구멍이 있다.'고 하니까."라고 그가 말했다. (O)	44 개
문장부호에 근거한 어절 오류	3. 외무부측은 "같은 차원의 교류 제의는 아니다."고 덧붙였다. (X) → 외무부측은 "같은 차원의 교류 제의는 아니다."라고 덧붙였다. (O) 4.아기를 안되, 조심해서 안아라. (O) 그러면 안되. (X) → 그러면 안되. (O)	7 개

[표 3] 문장부호 오류의 분류

IV.개선된 철자 검사기의 구현

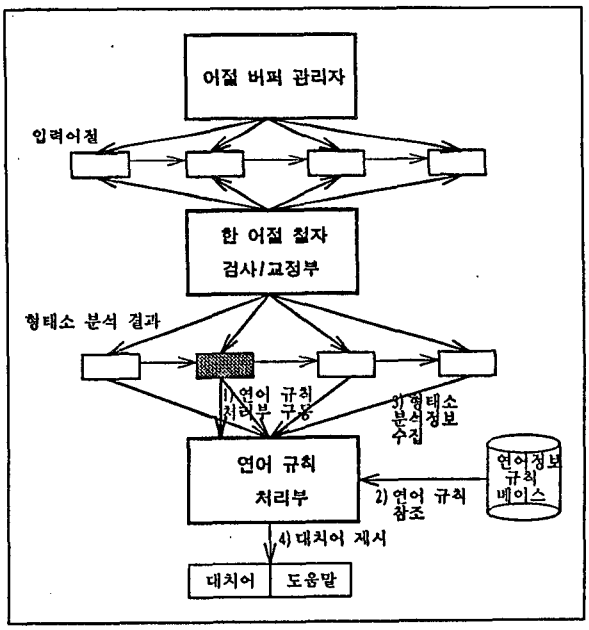


[그림 1] 개선된 철자 검사기 구성도

연어 정보에 기반한 개선된 철자 검사기는 1) 한 어절 단위 철자 검사 교정부, 2) 연어 규칙 처리부, 3) 문장 부호 처리부로 구성된다. 그리고 다수 어절을 동시에 접근하기 위해서 버퍼 관리자를 둔다. [그림 1]은 전체적인 시스템 구성도이다.

어절 버퍼 관리자는 환형큐(circular queue)로 구성되어 있으며, 어절 버퍼 관리자는 입력 문서로부터 다수 어절을 입력 받아 각 처리부에서 발생하는 요구에 따라 한 어절씩 제공해준다. 한 어절 철자 검사 교정부는 기존의 한 어절 단위의 철자 검사 교정기에 연어 규칙 처리부나 문장 부호 규칙 처리부를 위한 형태소 분석 정보를 제공하는 기능이 추가된 것이다. 한 어절 철자 검사기에서 제공하는 형태소 분석 정보는 다음과 같이 표현할 수 있다.

<형태소 분석 정보> = <분석경로>|<분석경로>*
<분석 경로> = <형태소 정보>|<형태소 정보>*
<형태소 정보> = <음절 길이, 상태 정보, 형태소 종류>



[그림 2] 연어 규칙 처리 과정

[그림 2]는 연어 처리부의 수행 과정을 도식화한 것이다. 모든 검사 어절에 대해 연어 규칙을 검증하는 것은 시스템의 성능을 급격히 저하시킬 위험이 있다. 그러므로 개선된 철자 검사기에서는 다음의 단계로 처리한다.

1) 한 어절 단위 철자 검사 교정부에서 연어 규칙 처리부를 구동시킨다. 이 때 철자 검사 교정부에서는 해당 단어의 연어 규칙 번호를 연어 규칙 처리부로 넘긴다.
2) 구동된 연어 규칙 처리부는 연어 규칙 번호로써 연어 지식 베이스를 탐색하여 규칙을 획득한다.
3) 규칙에 따라 필요한 형태소 분석 정보들을 수집한다.
4) 수집된 형태소 정보에 연어 관계 규칙을 적용하여 오류의 경우는 적절한 대치어와 도움말 정보를 제시한다.

한 어절 철자 검사 교정부에서 연어 규칙 처리부를 구동할 때 한국어의 지배-의존 관계에 기반한다. 즉, 지배소를 만나면 연어 규칙 처리부를 구동시킨다. 다음 [표 4]는 일반적인 한국어의 지배-의존 관계 규칙이다[8].

관계	지배소	의존소
수식	명사	관형사, 명사, 어미, 조사
격부여	조사	명사, 대명사, 수사, 조사, 부사
양상부여	어미	동사 어간, 형용사 어간, 어미
부가	형용사 어간	조사,명사,부사
부가	동사 어간	조사,명사,부사,어미
강조	부사	부사

[표 4] 한국어 지배 관계 규칙

문장 부호 처리부 역시 연어 규칙 처리부와 같이 한 어절 철자 검사 교정부에서 구동시키며, 구동된 문장 부호 처리부는 자신이 필요로 하는 정보를 수집하여 대치어와 도움말을 출력해준다.

추가의 분석이 필요한 경우만 좌우의 형태소 분석 정보를 수집하므로 연어 규칙 처리부와 문장 부호 규칙 처리부에 의한 속도 저하는 전체 시스템의 성능에 큰 영향을 미치지 않는다.

V.결론

본 논문에서는 한 어절 철자 검사기의 처리 범위를 다수 어절로 확장함으로써 문맥상 잘못된 어휘 사용과 문장 부호의 오류를 검사하고 교정하는 기법을 제안하고 있다. 의미적 오류를 교정하기 위해 사용자들이 많이 범하는 의

미적 오류의 유형을 분석하고 형태소 간의 연어 정보를 이용하여 오류를 처리하였다. 그리고 문장 부호 처리의 오류 및 문장부호에 근거하여 좌우의 오류 어절을 처리하는 기능을 구현하였다.

지식 베이스의 크기가 확장됨에 따라 지식 베이스의 탐색시간이 증가할 것이며, 현재의 구 단위에서 한 문장까지인 분석 범위가 다수 문장으로 확대되면 그 분석 시간도 문제시될 것이다. 이 문제에 대한 해결 방안이 추가로 연구되어야 한다.

본 논문에서 제시한 개선된 철자 검사기는 한국어 문체 검사기(Korean Style Checker)로서 의의를 가지며, 형태소 간의 연어 정보는 향후 파싱 등의 문장 분석이나 의미 분석에 중요한 자료로 이용될 수 있을 것으로 기대된다.

참고 문헌

[1] 이병훈, 윤준태, 송만석, "말 뭉치를 기반으로 한 한국어 철자 교정기의 구현", 한글 및 한국어 정보 처리 학술발표논문집, pp.285-293, 1993.

[2] 정한민, 이근배, 이종혁, "자판 특성을 이용한 Neuro-Fuzzy 한국어 철자 교정기의 구현", 한글 및 한국어 정보 처리 학술발표논문집, pp.317-328, 1993.

[3] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기", 부산대학교 전자계산학과 석사학위 논문, 1995.

[4] 미승우, 새 맞춤법과 교정의 실제, 어문각, 1988.

[5] 한글 연구회, 새 한글 맞춤법 및 용례집, 이사야, 1988.

[6] 국립국어연구원, 국어 순화 자료집, 1992.

[7] Chul-Min Sim, Min-Jung Kim, Hyuk-Chul Kwon, "Automatic Revision of Korean Texts by Collocation Words," pp.280-284, 1994.

[8] Hyck-Chul Kwon, Aesun Yoon, "Unification-Based Dependency Parsing of Governor-Final Languages," Proc. of Second International Workshop on Parsing Technology, Cancun, Mexico, pp.182-192, 1991

[9] 이영식, "사전 근사탐색과 Hueristics를 이용한 한국어 철자 오류 교정 시스템 구현",부산대학교 전자계산학과 석사학위 논문, 1994.

[10] 이종현, 오상현, "N-GRAM 한글 사전을 이용한 오인식 단어의 교정 알고리즘", 한글 및 한국어 정보처리 학술발표논문집, pp.271-283, 1993.

[11] 강재우, "접속 정보를 이용한 한국어 철자 띄어 쓰기 검사기의 설계 및 구현", 한국 과학 기술원 전산학과 석사학위 논문, 1990.

[12] 강승식, 이호석, 문유진, 김영택, "한국어 문법 검사/교정 시스템의 설계", 90 춘계 논문집, 17권 1호, 한국정보과학회, 1990.

[13] R.L.Kashyap, B.J.Oommen, "Spelling correction using probabilistic methods," Pattern Recognition Letters, pp.147-154, 1984.

[14] Thomas-N.Turba, "Checking for Spelling Typo graphical Errors in Computer Based Text," SIGPLAN Notices, pp.298-312, 1981.

[15] Masaki YAMASHINA, "Collocaltion Analysis in Japanese Text Input," Proceeding of COLING budapest , 1988.

[부록] 연어 관계 지식 베이스의 예 (본다/붓다/붙다)

```
1021

((본다
   (COL   (N, "쌀"), (N, "떡쌀"),(N, "강물"),(N, "재산")
          (P, "-이 본다"))
   (ACOL  (P, ("본어 않다", "불어 않다"))
          (P, ("본어 다니다", "불어 다니다"))))
(붓다
   (COL   ((C, 신체),
          (P, "-이 붓다"))
          ((C, 액체), (N, "씨앗"), (N, "불입금"), (N, "잿돈"), (N, "적금")
          (P, "-을 붓다")))
   (ACOL  (P, ("붓어 않다", "불어 않다"))
          (P, ("붓어 다니다", "불어 다니다"))))
(붙다
   (COL   (N, "불"), (N, "습관"), (N, "수당"), (N, "실력"), (N, "사람"), (N, "시험"),
          (N, "박")
          (P, "-이 붙다")
          (P, "-에 붙다"))
   (ACOL (P, ("붙어 만들다", "부어 만들다"))
          (P, ("붙어 만들다", "불여 만들다")))))


                1021 : 규칙 번호
                COL : 연어 관계(Collocation)
                ACOL : 연어 오류 관계(Anti-collocation)
                N : 명사(Noun)
                C : 분류 정보(Category)
                P : 문장 형식(Pattern)
```